

Modélisation de répartitions d'occurrences spatio-temporelles et épidémiologie végétale

Florencia Chimard¹, Jean Vaillant¹, Jean-Heinrich Daugrois²

¹EA 4540 LAMIA, Université des Antilles et de la Guyane
BP 592 Campus de Fouillole, 97157 Pointe-à-Pitre Cedex, Guadeloupe

²CIRAD, station de Roujol
Petit-Bourg, Guadeloupe
fchimard@univ-aq.fr

MOTS - CLÉS

Processus ponctuel
spatio-temporel
Loi a priori Stick-
breaking
Processus de
Dirichlet
Modèles bayésiens
hiérarchiques
Virus de la feuille
jaune
Mélange de lois
Canne à sucre

RÉSUMÉ

Les processus ponctuels temporel, spatial et spatio-temporel sont utilisés depuis un certain nombre d'années dans des domaines telles que la biologie, l'épidémiologie, la sismologie, la neurologie afin de modéliser des séries d'occurrences d'événements. Notre étude concerne l'apport des modèles de mélange de processus ponctuels dans de telles modélisations. Elle fait également apparaître l'utilité de l'approche bayésienne pour apprécier la variabilité associée à certains facteurs environnementaux. Le lien avec les lois a priori *stick-breaking* telles que le processus de Dirichlet est présenté. Une illustration est fournie avec l'étude de la propagation de la maladie de la feuille jaune de la canne à sucre. La répartition spatio-temporelle des cannes à sucre infectées par le Sugar Cane Yellow Leaf Virus (SCYLV) est étudiée dans une parcelle, saine lors de la plantation, située en Guadeloupe. Notre objectif principal est de caractériser puis de modéliser la dissémination du SCYLV au sein de cette parcelle.

KEY WORDS

Spatio-temporal
Point Process
Stick-Breaking Priors
Dirichlet Process
Hierarchical Bayesian
Models
Yellow Leaf Virus
Distribution mixture
Sugar cane

ABSTRACT

Modeling spatio-temporal occurrences distributions and vegetal epidemiology

Temporal, spatial and spatio-temporal point processes have been used for several years as a tool for modeling series of events in many fields such as biology, epidemiology, seismology, neurology. Our concern in this study is to discuss mixture models of point processes in a Bayesian framework aimed to explore environmental variability. Stick-breaking prior distributions like the Dirichlet process are presented. An illustration is given by the sugarcane yellow leaf disease propagation. The spatio-temporal distribution of sugar canes infected by the Sugar Cane Yellow Leaf Virus (SCYLV) is studied in a virus free plot at plantation. Our goal is to characterize and model the SCYLV propagation within this experimental plot.

1. Introduction

Beaucoup de phénomènes sont de nature stochastique et nécessitent pour être modélisés et prédits des techniques statistiques appropriées. Un processus ponctuel (Daley et Vere-Jones, 2003 ; van Lieshout, 2000) est un mécanisme stochastique qui modélise

des localisations de points dans un espace donné. Une répartition de points dans l'espace ou dans le temps peut-être considérée comme la réalisation d'un processus ponctuel (Vaillant, 1991 ; Vaillant, 1992). Elle se présente sous la forme d'un ensemble de coordonnées ou de dates d'occurrences. En général, pour faire de l'inférence statistique sur un processus, on ne dispose que d'une réalisation de ce processus (Vaillant *et al.*, 1997).

Les processus ponctuels temporel, spatial et spatio-temporel sont utilisés depuis un certain nombre d'années en biologie, épidémiologie, sismologie, neurologie, afin de modéliser des séries d'occurrences d'événements. Par exemple, en sismologie, on utilise souvent la théorie des processus ponctuels (Ogata, 1999 ; Holden *et al.*, 2003 ; Zhuang, 2006) pour analyser la répartition des tremblements de terre selon leur localisation, leur date ou encore leur magnitude. Si l'on s'en tient à la distribution des épicentres, qui peut-être représentée dans l'espace et dans le temps, on a un processus ponctuel spatio-temporel sans marque. On peut citer également la modélisation visant à étudier l'incidence de maladie (Diggle, 2005 ; Kottas *et al.*, 2007). Le processus de Poisson est souvent un modèle proposé pour l'étude des phénomènes homogènes dans le temps et/ou l'espace avec absence de corrélation.

Cependant, il y a beaucoup de phénomènes pour lesquels l'intensité varie dans le temps et/ou l'espace et présentant de l'autocorrélation (Vaillant, 1992). Des modèles alternatifs sont donc proposés en tenant compte des caractéristiques du phénomène étudié : certains auteurs étudient des processus de Cox (Brémaud, 1981 ; Brix et Diggle, 2001 ; Kottas et Sanso, 2007), d'autres des processus à mémoire tels que le processus d'Hawkes (Hawkes, 1974). Sous des conditions de régularité usuelle, la loi d'un processus ponctuel est complètement caractérisée par son processus intensité conditionnelle ($\lambda(\cdot)$). La modélisation se fait donc uniquement à travers ce dernier.

Dans le cas de séries temporelles d'événements se produisant en des zones géographiques, les processus ponctuels spatio-temporels et les lois de dénombrement associés sont tout à fait adaptés à la description, à l'explication éventuelle et à la prévision du phénomène étudié (Brix et Diggle, 2001 ; Diggle, 2005). L'intégration dans le modèle de processus ponctuels des connaissances *a priori* sur des variables de l'environnement, conduit à une approche bayésienne (Robert, 2005). Cette démarche subjective des probabilités a été l'objet d'un vif débat entre bayésiens et non bayésiens jusqu'aux années 80 où la puissance de calcul a permis l'éclosion d'outils très performants pour l'analyse bayésienne.

La méthodologie inférentielle bayésienne peut-être appliquée pour prendre en compte des effets présumés des variables environnementales à partir d'un modèle hiérarchique englobant les informations disponibles. Nous illustrons cette possibilité sur des données de propagation de la maladie de la feuille jaune de la canne à sucre (SCYLV) (Edon-Jock, 2008 ; Jacquet, 2008) provenant du CIRAD¹ de

Guadeloupe. La maladie de la feuille jaune est une maladie vectorielle émergente pouvant induire des pertes de rendements variables. L'analyse des risques sur la dissémination du pathogène, sa diversité, sa virulence et l'évolution de ses populations, ainsi que le comportement de la plante, est donc importante pour la production cannière. La répartition spatio-temporelle de l'infection des cannes à sucre par le SCYLV a été étudiée pour une parcelle plantée en Guadeloupe avec environ 1600 vitro plants sains. Un échantillonnage exhaustif des cannes à sucre a été effectué sur la parcelle. Pour chacun des plants de la parcelle un diagnostic par immuno-empreinte a été réalisé à différentes dates d'échantillonnage.

Dans ce qui suit, nous présentons quelques éléments concernant la modélisation par mélanges et l'approche bayésienne. Nous y introduisons aussi le processus de Dirichlet. La maladie de la feuille jaune, son agent causal ainsi que le dispositif expérimental utilisé sont ensuite détaillés. Les analyses statistiques effectuées et la modélisation hiérarchique que nous proposons sont alors exposées.

2. Modèle de mélanges pour répartition spatio-temporelle

2.1 Mélange de lois

Une façon de décrire des observations issues d'une population statistique comportant plusieurs catégories est de considérer un mélange de lois de probabilité. Le nombre de catégories peut-être infini dénombrable. On peut aussi faire intervenir une variable latente θ dont la valeur diffère d'une catégorie à l'autre. Soit p_i la proportion d'individus dans la *i*ème catégorie et f_i la densité de probabilité de la *i*ème catégorie pour le caractère étudié. Le mélange s'écrit alors :

$$f(\cdot) = \sum_{i=1}^{\infty} p_i f_i(\cdot | \theta_{c_i}) \quad (1)$$

où θ_{c_i} détermine la catégorie c_i avec $p_i = P(\theta = \theta_{c_i})$.

Une généralisation à des mélangeurs pour espace d'états Θ infini non dénombrable pour la variable θ est ainsi possible en notant G la loi de θ . C'est alors la loi mélangeante. Si on note $k(\cdot | \theta)$ la loi mélangée de paramètre θ (loi conditionnelle à θ), on a alors :

$$f(\cdot) = \int_{\Theta} k(\cdot | \theta) dG(\theta)$$

Loi mélangée (Mixed law)
Mélange (Mixture) Loi mélangeante (Mixing law)

Figure 1. Relation de mélange par rapport à θ

¹ Centre de coopération Internationale en Recherche Agronomique pour le Développement

2.2 Choix bayésien

On considère un ensemble de données $y = \{y_1, y_2, \dots, y_m\}$ pour lequel la réalisation y_i est issue d'une loi de probabilité de densité $f(\cdot|\theta_i)$, où le paramètre θ_i est inconnu et la densité f est connue.

La statistique bayésienne (Robert, 2005) suppose qu'il est possible d'avoir une intuition sur θ sans pour autant faire appel aux données y . Ces intuitions peuvent alors être traduites en termes de loi de probabilité, ce sont les lois *a priori*. La contribution principale apportée par un modèle statistique bayésien est donc de considérer une distribution aléatoire pour les paramètres. L'usage d'une loi *a priori* est vraiment novateur puisque l'on passe de la notion de paramètre inconnu à celle de paramètre aléatoire. Ces lois *a priori* peuvent être définies à partir de paramètres appelés par convention hyperparamètres. Dans la pratique, ces hyperparamètres sont supposés connus. On voit ainsi que nous pouvons inclure plusieurs niveaux d'*a priori*. Un modèle est dit hiérarchique s'il y a plusieurs niveaux de connaissances *a priori* sur les paramètres intervenant dans le modèle. Aussi, une représentation de cette hiérarchie est faite à travers le graphe bayésien dit graphe orienté acyclique qui permet de mieux visualiser les dépendances stochastiques (Ishwaran *et al.*, 2000). Ici, la notion de hiérarchisation n'est pas au sens géographique pour lequel il s'agit d'emboîtement de différentes échelles spatiales.

Toute la théorie bayésienne repose donc sur une formule simple permettant de passer d'une loi *a priori* sur les paramètres à une loi dite *a posteriori* sur les paramètres sachant les données. En résumé :

la distribution jointe (ou conjointe) de (θ, x) :

$$f(x, \theta) = f(x|\theta) \times \pi(\theta) ;$$

la densité marginale de x : $f(x) = \int f(x, \theta) d\theta$;

la loi *a posteriori* de θ , obtenue par la formule de Bayes :

$$\pi(\theta|x) = \frac{f(x|\theta) \times \pi(\theta)}{f(x)} \text{ si } f(x) \neq 0.$$

2.3 Modèles de mélange dans le cadre bayésien

Dans un cadre bayésien, la loi mélangeante G (voir figure 1) est parfois aléatoire de distribution *a priori* P . Soit y une réalisation d'un mélange donné par l'équation. On peut reformuler le modèle sous la forme hiérarchique suivante :

$$\begin{aligned} y | \theta_i &\sim f(\cdot|\theta_i) \\ \theta_i | G &\sim G \\ G &\sim P \end{aligned} \quad (2)$$

Plusieurs distributions *a priori* pour G sont envisageables dans le modèle (2). Dans le cadre de l'estimation de densité, nous nous intéressons au processus de Dirichlet (DP). Notons que ses extensions principales à savoir les processus Stick-Breaking (Ishwaran et James, 2001 ; Duan *et al.*, 2007) et les processus Stick-Breaking à noyaux (Dunson et Park, 2008) sont de plus en plus employés dans la littérature. On obtient donc :

$$\begin{aligned} y | \theta_i &\sim f(\cdot|\theta_i) \\ \theta_i | G &\sim G \\ G &\sim DP(\alpha, G_0) \end{aligned} \quad (3)$$

où $DP(\alpha, G_0)$ est le processus de Dirichlet de paramètres α et G_0 (voir § 2.4). La figure 2 nous montre le graphe acyclique orienté du modèle hiérarchique (3).

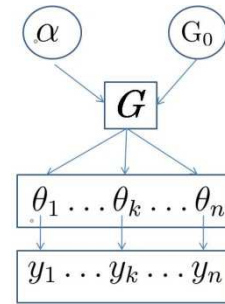


Figure 2. Graphe acyclique orienté du modèle(3)

Dans un cadre paramétrique, on suppose que la distribution aléatoire G (et donc f) est caractérisée par un paramètre $\phi \in \Phi$ inconnu de dimension finie. La densité aléatoire f appartient donc à l'espace de fonctions de dimension finie. L'*a priori* est donc défini sur ϕ et le modèle hiérarchique peut être reformulé de la façon suivante:

$$\begin{aligned} y | \theta_i &\sim f(\cdot|\theta_i) \\ \theta_i | \phi &\sim G(\cdot|\phi) \\ \phi &\sim p \end{aligned} \quad (4)$$

2.4 Processus de Dirichlet

En 1973 et 1974, Ferguson a publié des travaux sur la statistique bayésienne non paramétrique. Il introduit la notion de processus de Dirichlet (DP) comme mesure stochastique de probabilité. Plus précisément, considérons une distribution de probabilité G_0 sur un espace mesurable (X, H) . Une distribution G sur (X, H) est dite distribuée selon un DP si pour toute partition mesurable B_1, \dots, B_k de X , le vecteur des probabilités $(G(B_1), \dots, G(B_k))$ suit une distribution de Dirichlet associée à α et G_0 de la façon suivante :

$$(G(B_1), \dots, G(B_k)) \sim \text{Dir}(\alpha G_0(B_1), \dots, \alpha G_0(B_k)).$$

On note simplement : $G \sim DP(\alpha, G_0)$. Les DP sont alors définis à partir de deux paramètres. G_0 est appelée mesure de base ou espérance de G car pour tout borélien B , on a : $E(G(B)) = G_0(B)$. Le scalaire $\alpha > 0$ est le paramètre de précision (ou degré de confiance) et caractérise la variabilité de G autour de G_0 de la façon suivante : soient A et B deux boréliens

tels que $A \cap B = \emptyset$, alors :

$$\text{Var}(G(B)) = (G_0(B) \times (1 - G_0(B))) / (\alpha + 1)$$

$$\text{et : } \text{cov}(G(A), G(B)) = - G_0(A) \cdot G_0(B) / (\alpha + 1).$$

Si $G \sim DP(\alpha, G_0)$, alors la représentation stick-breaking de G introduite par Sethuraman en 1994 est de la forme suivante :

$$G(\cdot) = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}(\cdot),$$

$$p_k = V_k \prod_{j=1}^{k-1} (1 - V_j), \quad (5)$$

$$V_k \stackrel{i.i.d}{\sim} \text{Beta}(1, \alpha),$$

$$\theta_k \stackrel{i.i.d}{\sim} G_0$$

où $\delta_{\theta}(A)$ représente la mesure de Dirac en θ et vaut 1 si A contient θ et 0 sinon. Il est important de noter que la suite $p = (p_k)_{k=1, \dots}$ construite dans l'équation (5) satisfait que les p_k sont des variables aléatoires, appelées poids aléatoires, indépendantes de θ_k tel que $0 \leq p_k \leq 1$ et $\sum_{k=1}^{\infty} p_k = 1$. On voit que G_0 détermine les positions des composantes discrètes de G , tandis que α règle la variance des poids p_k .

Les réalisations d'un processus de Dirichlet peuvent donc être vues comme un mélange d'un nombre infini dénombrable de lois.

Loi a posteriori

Une motivation fondamentale pour l'utilisation du processus de Dirichlet comme mesure stochastique de probabilité est la simplicité de la mise à jour de la loi *a posteriori* qui est également un processus de Dirichlet (Teh et al., 2004). Considérons le modèle suivant :

$$y_1, y_2, \dots, y_n | G \stackrel{i.i.d}{\sim} G$$

$$G | \alpha, G_0 \sim DP(\alpha, G_0) \quad (6)$$

Alors la loi a posteriori de G est :

$$G | y_1, \dots, y_n \sim DP(\alpha^*, G_0^*)$$

$$\text{où } \alpha^* = \alpha + n \text{ et } G_0^* = \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^n \delta_{y_i}.$$

La loi prédictive, obtenue en marginalisant selon la loi G , admet la représentation en urne de Polya suivante :

$$y_{n+1} | y_1, \dots, y_n \sim \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{j=1}^n \delta_{y_j}. \quad (7)$$

On peut déduire directement de l'équation (7) que conditionnellement aux variables latentes y_1, \dots, y_n qui sont déjà échantillonnées, la probabilité que la prochaine valeur soit l'une des valeurs précédemment obtenues est $\frac{1}{\alpha + n}$, alors que la probabilité d'une toute nouvelle valeur est $\frac{\alpha}{\alpha + n}$.

3. La maladie de la feuille jaune de la canne à sucre

3.1 La canne à sucre

Le terme canne à sucre désigne un ensemble d'espèces de plantes de la famille des Poaceae et du genre *Saccharum*². Elles sont cultivées pour leurs tiges, dont on extrait du sucre. Avec un volume annuel de production supérieur à 1,3 milliard de tonnes, ce sont les premières plantes cultivées au plan mondial avec près de 23% de la masse totale produite en agriculture dans le monde. Elles furent jusqu'au début du XIX^e siècle la seule source importante de sucre et représentent toujours actuellement 65 à 70% de la production de sucre.

En France, la canne à sucre est cultivée dans plusieurs départements d'outremer : essentiellement en Guadeloupe et en Martinique, du côté des Antilles, et à l'île de la Réunion, dans l'océan Indien.

La culture de la canne à sucre demeure l'un des fondements de l'économie et même de la société guadeloupéenne. La sole cannière est prépondérante dans le nord-est de la Basse-Terre. Mais elle subit depuis une vingtaine d'années un contexte international défavorable et ce, malgré plusieurs plans de relance. La production de sucre à des coûts compétitifs demeure le principal enjeu actuel de la culture de la canne à sucre malgré la diversification des usages de la canne à sucre (énergie, éthanol, rhum). Les cultures de diversification (cultures maraîchères et fruitières) sont de plus en plus représentées, mais ne parviennent pas à s'imposer comme culture de remplacement de la canne à sucre.

Les maladies et les ennemis de la canne sont très nombreux et ont souvent un impact important sur le rendement des cultures. Il peut s'agir d'insectes et de rongeurs, mais aussi de champignons, de bactéries et de virus. Ils infligent parfois d'importants dégâts et peuvent provoquer des pertes de rendements significatives. Beaucoup d'efforts de recherche sont donc investis pour lutter contre les ennemis de la canne. Il existe plus d'une soixantaine de maladies de la canne causées soit par champignons, des bactéries ou des virus. Parmi les maladies virales, on trouve la maladie de la feuille jaune.

² <http://fr.wikipedia.org/wiki/Plante>

3.2 Le virus SCYLV

La maladie de la feuille jaune s'exprime par des symptômes de jaunissements de la partie centrale du limbe alors que le reste du limbe demeure vert. Cette coloration peut également être rose (figure 3). L'agent causal, le SCYLV est un polerovirus, famille des luteoviridae. Il envahit les cellules du phloème de la plante. Ce virus peut être transmis par plusieurs espèces de pucerons vecteurs.



Figure 3. Symptômes de la maladie de la feuille jaune observés sur une parcelle en Guadeloupe
J.-H. Daugrois (CIRAD)

Durant toutes les études menées aux Antilles, le seul vecteur du SCYLV observé est le *M. sacchari*. C'est un puceron de un à deux millimètres de long qui existe sous forme d'aptère ou d'ailé.

La présence du virus dans les plantes a été diagnostiquée pour cette étude par immuno-empreinte de tissus foliaire. Brièvement le diagnostic consiste à prélever la première feuille entièrement développée, d'effectuer dans la partie basse de la feuille une section nette transversale de la partie centrale après avoir ôté les parties externes du limbe. La section transversale est ensuite imprimée sur une membrane de nitrocellulose par une pression rapide. L'empreinte ainsi obtenue est ensuite révélée par une double réaction antigène/anticorps (Edon-Jock, 2008).

3.3 Données observées

Une parcelle de cannes à sucre d'une superficie d'environ 1300 m² a été mise en place pour l'analyse de la transmission spatio-temporelle du SCYLV par son vecteur. Sur cette parcelle, environ 1600 vitro plants sains ont été plantés. Les plantes sont disposées sur des rangs distants de 1,5 m et la distance des plantes dans le rang est de 0,5 m. Pour l'étude de la répartition spatio-temporelle des plants infestés par les pucerons dans la parcelle, un échantillonnage exhaustif des cannes à sucre a été effectué sur la parcelle. Ceci impliquant l'observation de tous les plants afin de minimiser le risque d'erreur.

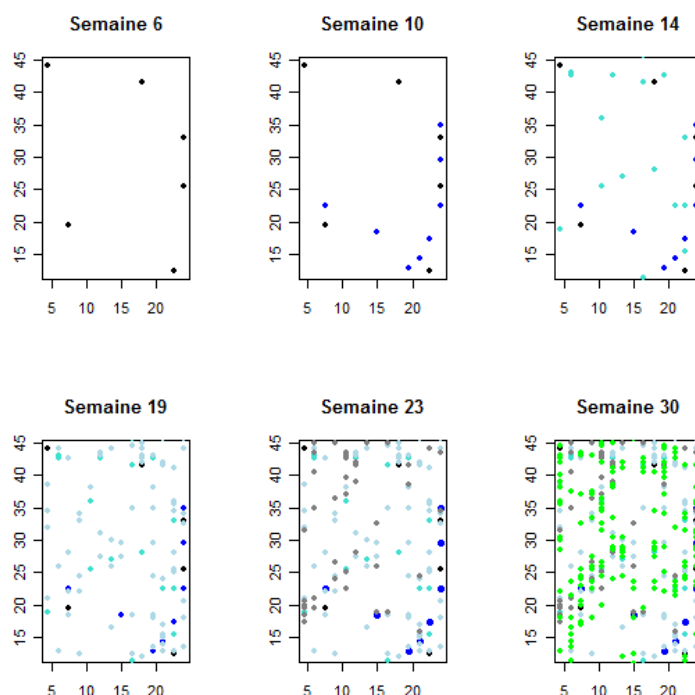
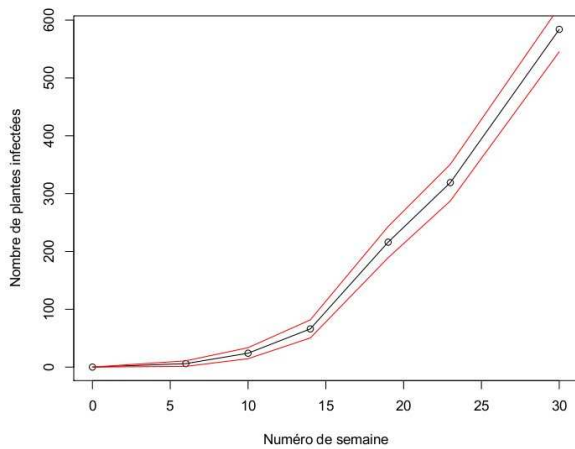


Figure 4. Carte de l'évolution spatio-temporelle des infections dans la parcelle par le SCYLV

Chacun des plants de la parcelle a été diagnostiqué quand à la présence du SCYLV par immuno-empainte à 6 dates d'échantillonnage. La parcelle représente donc la population statistique et les cannes à sucre les individus statistiques. À partir de ce diagnostic, des cartes d'infections représentant la répartition spatio-temporelle des plantes infectées à chaque date d'échantillonnage ont été obtenues (figure 4).

Les différentes couleurs utilisées traduisent les dates de première détection de la contamination d'une plante. Nous observons que le niveau d'infection augmente progressivement de la semaine 6 jusqu'à la semaine 30.



--- zone de confiance de sécurité 95%

Figure 5. Evolution temporelle du nombre de plantes infectées sur la parcelle observée

4. Modélisation hiérarchique

Chaque point (dans la réalisation du processus ponctuel) représente ici une plante infectée. Nous disposons de la position exacte de chaque plante infectée sur la parcelle et par conséquent, la position de chaque nouvelle infection. Si nous considérons x la position spatiale d'une plante quelconque de la zone étudiée X et $y_{x,t}$ son état sanitaire à la date t , la loi des observations $(y_{x,t})_{x \in X}$ doit prendre en compte la nature binaire des $y_{x,t}$ et le fait que X soit au plus dénombrable.

Ainsi, $y_{x,t}$ suit donc une loi de Bernoulli de paramètre noté $p_{x,t}$ qui n'est autre que la probabilité d'être infecté à la date t et à la position x . Cette probabilité peut-être fixe pour une approche classique, cependant il arrive souvent que les conditions environnementales varient de façon conséquente. Il paraît naturel d'adopter une approche bayésienne qui peut-être présentée

comme une généralisation de l'approche classique : le paramètre $p_{x,t}$ n'est plus une valeur fixe inconnue mais une variable aléatoire dont il faut spécifier la loi dite *a priori* au moment de la modélisation.

Dans les données traitées par Jacquet (2008) et Edon-Jock (2008), deux phases ont été distinguées dans la contamination de la parcelle saine. Une première phase où les contaminations semblent se faire selon une répartition complètement aléatoire et une seconde phase où on observe une agrégation des plantes infectées. Ceci est confirmé par différents tests statistiques basés sur la dispersion spatiale, l'autocorrélation spatiale et spatio-temporelle. La première phase correspondrait à l'arrivée des premiers aîlés sur la parcelle tandis que la seconde phase coïncide à la fermeture de la parcelle c'est-à-dire à un stade phénologique des plantes pour lequel les feuilles de cannes voisines se touchent. Cette dernière situation permet le déplacement des pucerons aptères d'une plante à l'autre favorisant ainsi la contamination de proche en proche à partir des infections primaires.

Compte tenu de ces informations, nous formalisons un modèle en distinguant les deux étapes décrites ci-dessus. Nous utilisons les notations suivantes :

τ : la date de changement de phase,

a_t : le paramètre d'intensité d'arrière plan à la date t ,

θ : le paramètre d'interaction,

f_θ : la fonction de contact qui dépend de la distance $x-y$ entre deux individus,

H_{t-} : histoire de la parcelle jusqu'à la date t exclue.

$$\lambda(x, t | H_{t-}) = \begin{cases} a_t \times 1_{[y_{x,t}=0]} & \text{si } t \leq \tau \\ \left(a_t + \sum_{y \in I_{t-}} f_\theta(x, y) \right) \times 1_{[y_{x,t}=0]} & \text{si } t > \tau \end{cases} \quad (8)$$

Ce modèle est une généralisation de celui proposé par Jacquet *et al.* (2009) car l'intensité d'arrière plan varie au cours du temps. D'autre part, la variabilité stochastique des paramètres du modèle est prise en compte et est introduite par le biais de lois *a priori* sur ces paramètres. Nous choisissons comme *loi a priori* pour le processus (a_t) un processus de Dirichlet. Nous pouvons reformuler le problème d'estimation de densité selon le modèle hiérarchique suivant :

$$y_{x,t} | p_{x,t} \sim \text{Bernoulli}(p_{x,t}) \quad (9)$$

$$(a_t) \sim DP(\alpha, G_0)$$

$$\text{où } p_{x,t} = 1 - e^{-\int_0^t \lambda(x,u) du}.$$

Un processus de Dirichlet comme *a priori* sur le processus temporel d'intensité d'arrière plan (a_t) signifie que l'on ne se restreint pas à une unique loi (de processus temporel) pour caractériser l'information *a priori* disponible mais à une famille de lois de processus temporel. Cette approche est dite non paramétrique ou parfois semi-paramétrique. En l'occurrence, il paraît naturel que cette famille soit centrée sur un processus log-normal à incréments indépendants (Moller *et al.*, 1998). L'inférence de tels modèles fondés sur le processus de Dirichlet est en général conduite à partir de méthodes de Monte Carlo par chaînes de Markov (MCMC) et plus particulièrement les algorithmes de Gibbs (Iswharan et James, 2001).

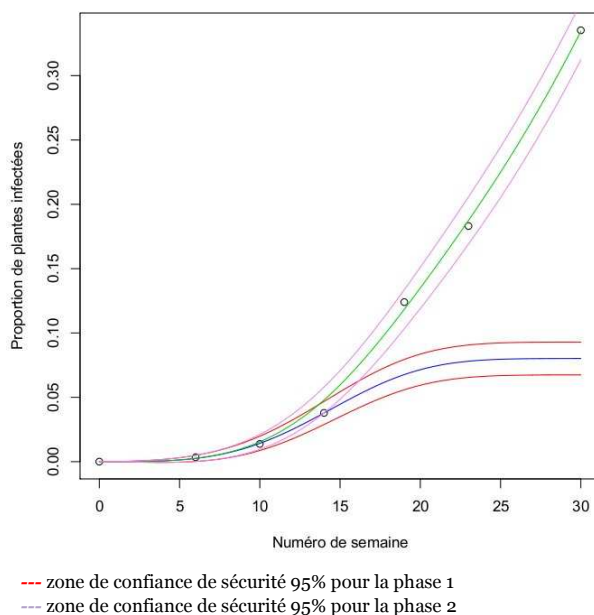


Figure 6. Evolution de la proportion de plantes infectées par le virus

La figure 6 nous montre les résultats des calculs effectués. Un changement de phase est constaté à la semaine 14. Pour chaque phase, nous avons concentré notre attention sur deux résumés numériques : d'une part, l'intensité moyenne d'arrière plan et d'autre part, un paramètre d'échelle lié à la fonction de contact et traduisant l'accentuation du niveau d'intensité due à la mobilité des pucerons aptères. Pour la phase 1, nous obtenons une intensité moyenne d'arrière plan de 0,084 et un paramètre d'échelle valant 16,041. Pour la seconde phase, qui semble démarrer entre les semaines 10 et 14 pour les données étudiées, l'intensité moyenne d'arrière plan vaut 3,986 et le paramètre d'échelle 60,410.

L'intensité moyenne globale serait la résultante d'un nombre aléatoire de composantes qui agissent sur le degré d'infection: variables environnementales, prédation, flux d'ailés, production d'aptères et degré de colonisation des plantes et de mobilité entre plantes voisines.

Une stratégie de lutte efficace contre la propagation de la maladie devra donc passer par une action de limitation des facteurs favorables à sa dissémination avant la colonisation et production parthénogénétique des pucerons aptères. La courbe représentative de la proportion attendue de plantes infectées devrait alors se ramener à la courbe prolongeant la phase 1.

5. Conclusion et discussion

Nous avons proposé un modèle à deux phases caractérisé par sa fonction d'intensité conditionnelle dont l'expression à l'instant t et au site x est la somme d'une série chronologique à termes positifs et d'une fonction dépendant des proximités entre x et les sites infectés avant la date t . Cette fonction de contact ne commence à dépendre des infections passées qu'à partir d'une date τ . À partir de la théorie des processus ponctuels et des propriétés de mélange de lois, une approche bayésienne est possible en incluant des lois *a priori* stick-breaking, dont le processus de Dirichlet est un exemple. On dispose alors d'outils de traitement statistique prenant en compte l'évolution stochastique du phénomène observé. Ainsi, dans le cas de l'état sanitaire des parcelles de cannes à sucre infectées par le virus SCYLV, la variabilité est mieux prise en compte en considérant des paramètres de phase spatio-temporels. Cette approche combinant la mise au point d'algorithmes MCMC hybrides peut être d'une grande utilité à condition d'avoir des données pertinentes pour étudier le phénomène d'intérêt.

Pour ce qui concerne la propagation du SCYLV sur les plantations de canne à sucre, il est important d'avoir des éléments de description et de modélisation de l'évolution spatio-temporelle de la maladie de la feuille jaune pour mettre au point une veille épidémiologique satisfaisante. Le modèle proposé est multi-phases à intensité conditionnelle dépendant d'une intensité d'arrière plan (a_t) stochastique et mérite d'être étudié et affiné, le cas échéant, sur d'autres dispositifs expérimentaux. Dans un article en cours, nous discutons d'une version en temps discret de ce modèle et de la construction d'algorithmes de type Metropolis-within-Gibbs pour l'échantillonnage de la distribution *a posteriori* des paramètres d'intérêt.

6. Références bibliographiques

- Brémaud P., 1981, *Point Processes and Queues, Martingales Dynamics*, Springer Verlag, New York.
- Brix A., Diggle P. J., 2001, Spatiotemporal prediction for log-gaussian Cox processes, *Journal of the Royal Statistical Society B*, 63, 823-841.
- Célini L., Vaillant J., 1999, Répartition spatio-temporelle des présences d'ailés d'Aphis gossypii (Glover) (Hemiptera:aphididae) en culture cotonnière (Malvaceae), *Canadian Entomologist*, 131, 813-824.
- Chimard F., Vaillant J., Estimation bayésienne de l'intensité d'arrière plan stochastique d'un processus spatio-temporel, En préparation.
- Daley D. J., Vere-Jones D., 2003, *An Introduction to the Theory of Point Processes*, Springer, New York.
- Diggle P. J., 2005, A partial likelihood for spatio-temporal point processes, *The Berkeley Electronic Press*, 75.
- Duan J. A., Guindani M., Gelfand A. E., 2007, Generalized spatial dirichlet process models, *Biometrika*, 94, 809-825.
- Dunson D. B., Park J.-H., 2008, Kernel Stick Breaking Processes, *Biometrika*, 95, 307-323.
- Edon-Jock C., 2008, Le virus de la feuille jaune de la canne à sucre : spécificités de la dissémination en Guadeloupe et en Martinique, Thèse de doctorat, Université des Antilles et de la Guyane.
- Ferguson T. S., 1973, A bayesian Analysis of Some Nonparametric Problems, *The Annals of Statistics*, 1, 209-230.
- Ferguson T.S., 1974, Prior Distributions on Spaces of Probability Measures, *The Annals of Statistics*, 2, 615-629.
- Hawkes A. G., Oakes D., 1974, A cluster representation of a self-exciting process, *Journal of Applied Probability*, 11, 493-503.
- Holden L., Sannan S., Bungun H., 2003, A stochastic marked point process model for earthquakes, *Natural Hazards and Earth System Sciences*, 3, 95-101.
- Ishwaran H., James L. F., 2001, Gibbs sampling methods for stick-breaking priors, *Journal of the American Statistical Association*, 96, 161-173.
- Ishwaran H., Zarepour M., 2000, Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models, *Biometrika*, 87, 371-390.
- Jacquet O., 2008, Analyse statistique des processus ponctuels spatio-temporels de propagation sur une grille, Thèse de doctorat, Université des Antilles et de la Guyane.
- Jacquet O., Vaillant J., Edon-Jock C., Daugrois J.-H., 2009, Statistical analysis of partially observed two-phase point processes, *Computational Statistics and Data Analysis*, soumis.
- Kottas A., Duan J.A., Gelfand A. E., 2007, Modeling disease incidence data with spatial and spatio-temporal dirichlet process mixtures, *Biometrical Journal*, 49, 1-14.
- Kottas A., Sanso B., 2007, Bayesian mixture modeling for spatial poisson process intensities, with applications to extreme value analysis, *Statistical Planning and Inference*, 137, 3151-3163.
- J. Møller, A. R. Syversveen, Waagepetersen R. P., 1998, Log Gaussian Cox processes, *Scandinavian Journal of Statistics*, 25, 451-482.
- Ogata Y., 1999, Seismicity analysis through point process modeling: a review, *Pure and Applied Geophysics*, 155, 471-507.
- Robert C., 2005, *Le choix bayésien, principes et pratique*, Springer, Paris.
- Sethuraman J., 1994, A constructive definition of dirichlet priors, *Statistica Sinica*, 4, 639-650.
- Vaillant J., 1991, Negative Binomial Distributions of Individuals and Spatio-Temporal Cox Processes, *Scandinavian Journal of Statistics*, 18, 235-248.
- Vaillant J., 1992, Echantillonnage et étude statistique de populations en milieu hétérogène, *Statistique Appliquée*, 4, 15-26.
- Vaillant J., Argente G., Boulard C., 1997, Hypodermose bovine: une procédure quasi séquentielle d'observation des cheptels pour la surveillance de la recrudescence en zone éradiquée, *Veterinary Research*, 28, 461-471.
- Van Lieshout M. N. M., 2000, *Markov Point Processes and their Applications*, Imperial College Press, London.
- Teh Y. W., Jordan M. I., Beal M. J., Blei D. M., 2004, Hierarchical dirichlet processes, *Journal of the American Statistical Association*, 101.
- Zhuang J., 2006, Second-order residual analysis of spatiotemporal point processes and applications in model evaluation, *Journal of the Royal Statistical Society*, 64, 635-653.